DENOISING COMPOSITION IN DISTRIBUTIONAL SEMANTICS

Gábor Borbély
Department of Algebra
Budapest University of Technology and Economics
borbely@math.bme.hu

András Kornai, Dávid Nemeskey
Institute for Computer Science
Hungarian Academy of Sciences
andras@kornai.com,
nemeskeyd@sztaki.mta.hu

Marcus Kracht
Department of Linguistics
Bielefeld University
marcus.kracht@uni-bielefeld.de

INTRODUCTION

In distributional semantics we derive the embedding from a corpus, and the corpus is just a sample from the entire distribution. We analyze the noise of the obtained vectors and other sources of noise, and how much the considerations of compositionality discussed in [4] are affected by noise.

NOISE EFFECTS

We took a standard English corpus, the UMBC Webbase [3], and a new Hungarian corpus of comparable size. We cut UMBC in two roughly equal parts in two ways: *even-odd* cut and *begin-end* cut. The Google analogy task (GA) was used to measure the quality of the linear structure.

We also trained the GloVe [5] vectors on a morphologically analyzed Hungarian corpus where the stem was treated as separate from the suffix (see 'GF'). After running separately GloVe on the odd and the even parts, we compared the cosine similarities of the vectors in the two embeddings by five different methods, and we repeated the experiment comparing the beginning and end halves of UMBC, and the even-odd cut on the Hungarian corpus.

The first (direct) comparison is the cosine similarity of the words in the different embeddings. We trained linear transformations to bring the vectors obtained from the two subcorpora into closer alignment. We do this by fitting the best orthonormal transformation (*rot*), or by the best general linear transformation (*gl*), without normalization for vector length (*nolen*) and with normalization (*len*). Column *@100* shows the results for the first 100 most frequent words; column *@5k* shows 100 less frequent words, between 4,900 and 5,000; and column *@50k* the average similarity of the first 50k words.

cut	cond	@100	@5k	@50k
even - odd GA 71.5%-71.7%	direct	.010	.004	.003
	nolen-rot	.973	.946	.863
	len-rot	.973	.945	.862
	nolen-gl	.977	.955	.880
	len-gl	.976	.952	.879
beg - end GA 71.8%-70.7%	direct	.002	.004	.003
	nolen-rot	.966	.898	.764
	len-rot	.966	.897	.763
	nolen-gl	.965	.908	.789
	len-gl	.964	.903	.787
Hun even - odd	direct	.357	.107	.072
	nolen-rot	.905	.884	.824
	len-rot	.903	.881	.823
	nolen-gl	.908	.899	.846
	len-gl	.903	.894	.844

'GF' CORPUS

To study stems and inflections separately, in laboratory pure form, we took a large corpus of a highly agglutinative language, Hungarian, and by morphological analysis produced a *de-glutinized* version where the stem and the paradigmatic suffixes are separated by a whitespace the same way two words would. Perhaps inevitably, this became known as the *gluten free* corpus of Hungarian.

az	emberi	méltóságért	és	békéért
a	emberi	méltóság CAU	és	béke CAU
		folytatott		harccal
		folytat[PERF_PART]		harc INS

Clusters of similar words naturally appear. Remarkably, postpositions such as *alatt* cluster not just with other postpositions but also with case endings: the nearest neighbors are the terminative, inessive, superssive cases, the postposition *után* 'after', the adessive case, the postposition *között* 'between', followed by the illative and sublative cases.

COMPOSITIONALITY

In [4] we used the compositional mechanism of Context Vector Grammars [6] to demonstrate that grammatical formatives such as the deadjectival adverb-forming suffix *-ly* or the comparative *-er* must contribute additively to the representations, so that e.g.

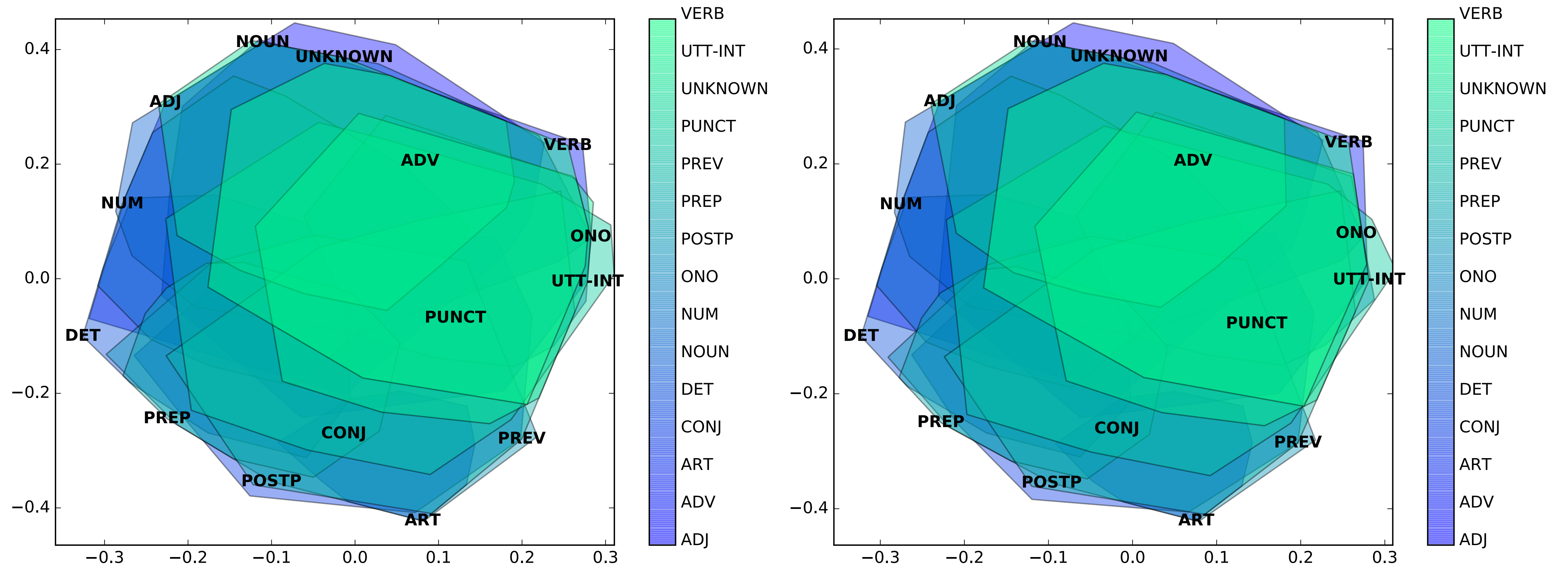
$$\vec{bigger} - \vec{big} + \vec{small} \approx \vec{smaller}. \quad (1)$$

The fact that embeddings perform well in 'grammatical' GA tasks such as `gram3-comparative` or `gram1-adjective-to-adverb` suggests that there exists e.g. a morpheme vector $\vec{ER}$ for comparative. With the GF embedding, such morphemes are part of the vocabulary and vectors are explicitly learned for them; (1) thus becomes the exact (and trivial) equation

$$(\vec{big} + \vec{ER}) - \vec{big} + \vec{small} = (\vec{small} + \vec{ER}). \quad (2)$$

LINEAR STRUCTURE

The POS clusters of GF Hungarian corpus, projected into two dimensions. The two figures correspond to the even and odd cut of the text.



HIGH DIMENSIONAL CONES

In d dimensions, the unit ball has surface area $2\pi^{d/2}/\Gamma(d/2)$. If we equally divide this area among n cones, each peaking at the origin and having half angle θ, the surface area cut out by one cone is equal to $1/n$th of the total surface:

$$\frac{2\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d-1}{2})} \int_0^\theta \sin^{d-2}\phi \, d\phi = \frac{1}{n} \cdot \frac{2\pi^{d/2}}{\Gamma(d/2)} \quad (3)$$

In the usual case when $n \sim 10^5, d = 300$, this gives us a noise cone of about 0.25. In practice, the winners in analogy task lookups often display cosine similarities in the 0.4–0.5 range, well above the noise level.

SPARSE OVERCOMPLETE

We considered sparse overcomplete representations computed of the same GloVe vectors by the method of [2][a]. The rotated and general linear similarities are shown, as well as GA results. The raw numbers are not directly comparable to those obtained for GloVe, since here $d = 3000$, for which (3) yields .08, about a third of the .25 we obtained in 300 dimensions. In this light, the sparse vectors are *more* stable than the raw GloVe vectors we obtained them from. The sparse vectors have 3-600 nonzero elements out of 3,000.

The nonnegative vectors were also considered in the same spirit as [2] and these are again quite stable.

We investigated the considerably more sparse vectors suggested by [1], reimplementing their method using the `pyksvd` library[b]. These vectors have at most 5 nonzero components out of 2,000, referred as 'k=5'.

vecs	dim	cond	@100	@5k	50k
Sparse 53.5%-53.7%	3k	nolen-rot	.627	.536	.458
	3k	nolen-gl	.754	.688	.600
Nonneg 44.2%-45.7%	3k	nolen-rot	.532	.477	.415
	3k	nolen-gl	.621	.599	.553
k=5 23.1%-23.5%	2k	nolen-rot	.523	.466	.505
	2k	nolen-gl	.583	.515	.561

REFERENCES

[1] S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski. Linear algebraic structure of word senses, with applications to polysemy. *arXiv:1601.03764v1*, 2016.

[2] M. Faruqui, J. Dodge, S. Jauhar, C. Dyer, E. Hovy, and N. Smith. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL 2015*, 2015. Best Student Paper Award.

[3] L. Han, A. L. Kashyap, T. Finin, J. Mayfield, and J. Weese. Umbc_ebiquity-core: Semantic textual similarity systems. In *Second Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 44–52, Atlanta, Georgia, USA, 2013. Association for Computational Linguistics.

[4] A. Kornai and M. Kracht. Lexical semantics and model theory: Together at last? In *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 14)*, pages 51–61, Chicago, IL, July 2015. Association for Computational Linguistics.

[5] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, 2014.

[6] R. Socher, J. Bauer, C. D. Manning, and N. Andrew Y. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 455–465, Sofia, Bulgaria, 2013. Association for Computational Linguistics.

[a] https://github.com/mfaruqui/sparse-coding

[b] https://github.com/hoytak/pyksvd

[c] Research partially supported by National Research, Development and Innovation Office NKFIH grant #115288.